The influence of semantic features on lexical geographical variation

Karlien Franco, Dirk Geeraerts, Dirk Speelman
QLVL, Department of Linguistics
KU Leuven
Leuven, Belgium
{karlien.franco, dirk.geeraerts, dirk.speelman}@arts.kuleuven.be

***Abstract*—In this paper, we investigate the influence of semantic concept features on lexical geographical variation. More specifically, we take an onomasiological approach to inquire into the effect of concept vagueness, salience, affect and semantic field. We use quantitative operationalizations of these features as predictors in a linear regression analysis. Our response variable is a composite variable that takes into account the number of variants per concept and the degree to which the concepts are scattered across geographical space in a heterogeneous way. Our model reveals that vaguer, less salient and non-neutral concepts show significantly more variation and that the lexical variants for these concepts are scattered across geographical space in a less homogeneous way. We also find differences between semantic fields.**

Keywords— dialectometry, lexical variation, Dutch, quantitative lexicology

I. BACKGROUND[1]

In dialectometry, language variation is often assumed to be governed by lectal or geographical factors [2]–[4]. However, a pilot study on the semantic field *the human body* showed that semantic features can influence lexical geographical variation as well [5], [6]. More specifically, the pilot study provided significant evidence for the influence of concept vagueness, salience and negative affect.

The selection of these features is, on the one hand, motivated by a prototype-theoretical view of language. Semasiological features that have been discussed in the context of prototype theory seem to influence onomasiological variation across dialects. Vaguer concepts (concepts with fuzzy boundaries that are not easily distinguished from related concepts), such as ACHTERSTE and ACHTERWERK ('bottom'), often show more lexical geographical variation than less vague concepts, such as DUIM ('thumb'). Less salient concepts (concepts that are psychologically less entrenched), such as SLUIK HAAR ('straight hair'), show more lexical geographical variation than more salient concepts like HOOFD ('head').

On the other hand, the pilot study also focused on a more traditional semantic feature, viz. the degree to which a concept has a negative connotation. The inclusion of negative affect as a factor of lexical geographical variation is inspired by the fact that taboo-laden concepts often show a high degree of lexical richness [7], [8]. This is also confirmed by the pilot study: negatively connoted concepts, such as KWIJL ('drool') show more lexical geographical variation than neutral concepts like JUKBEEN ('cheekbone').

In this paper, we elaborate on the results of the pilot study in two ways. First, we expand the scope to other semantic fields than *the human body*. As a result we are able to show that the influence of concept features on lexical geographical variation is relatively stable across different semantic fields. In addition, we can determine which differences occur in dialectal variation in different semantic fields. Methodologically, we also take into account alternative operationalizations of two predictors that were used in the pilot study, viz. salience and negative affect.

II. DATA

For our analysis, we use data that were collected in the Limburgish dialect area. The Limburgish dialect is a variety of Dutch, spoken in the Dutch and Belgian provinces of Limburg (located in the south of the Netherlands and in the east of the Dutch speaking part of Belgium). We use the digitized database of the *Woordenboek van de Limburgse Dialecten* ([9]; 'dictionary of Limburgish dialects'). This dictionary is based on three types of data. First, it contains data from large-scale questionnaires that were sent out across the dialect area. Some of these questionnaires, in particular the questionnaires that were distributed by the *Nijmeegse Centrale voor Dialect- en Naamkunde* (NCDN; 'center for dialectology and onomastics of Nijmegen'), contain questions and pictures that are specifically used to elicit data for the Limburgish dialect dictionary. Other questionnaires serve a more general purpose (like the data from the *Reeks Nederlandse Dialectatlassen;* 'series of Dutch dialect atlases'). Second, the dictionary includes data described in small-scale local dictionaries. For example, it contains materials from dictionaries like *Het Bjêvels* [10], a dictionary of the dialect of Beverlo (a small town in the west of the Belgian province of Limburg). Third, the dictionary of Limburgish dialects relies on other sources, such as student essays, dissertations about a particular dictionary and local journals like *Veldeke: tijdschrift voor*

[1] A more detailed overview of this study can be found in [1].

Limburgse volkscultuur (‘Veldeke, journal of Limburgish folklore’).

Since the operationalization of some of the predictor variables assumes that the data that we use were collected in a systematic way, i.e. with questionnaires that were *systematically* sent out across the *entire* dialect area, we only use part of the data base in our analysis. More specifically, we rely on the questionnaires that were sent out by the NCDN with the specific purpose of collecting dialect material for this particular dictionary. Most of these data were collected in the period 1960-1990. Follow-up surveys were distributed between 1997 and 2005.

The dictionary of the Limburgish dialects is an onomasiological dictionary. The concepts are organized into three large parts (farming terminology, non-agricultural specialist terminology and general vocabulary), which are further divided into chapters that each represent a particular semantic field. In this study, we rely on four digitally available chapters of the part on general vocabulary. To compare our results to the findings of the pilot study, we include the chapter relating to *the human body*. Additionally, we include the chapters about *the physical and abstract world*, about *personality and feelings*, and about *family and sexuality*.

In general, the dictionary contains a fairly large collection of dialect words from a relatively large amount of locations in Limburg for each of the concepts. For the concept BRUIDSJAPON (‘wedding dress’), for instance, the data set contains 93 entries collected in 51 different locations; for the concept WOENSDAG (‘Wednesday’), 308 dialect entries are available from 117 places in Limburg. However, as it turns out that some concepts only have responses for a small number of locations in the data base, we exclude concepts that occur in 50 or less places. We assume that these relatively small numbers point to a lack of consistency in the distribution of the questionnaire. We adopt the same assumption for locations that occur with only few concepts in the data base: places that are only represented by 50 or less concepts in the data base, are excluded from the analysis. Overall, the data set that is used for the analysis contains 859 concepts for 243 places. We investigate 180 concepts relating to *the human body*, 243 concepts from *the physical and abstract world*, 317 concepts relating to a person’s *personality and feelings* and 119 concepts from the semantic field of *family and sexuality*.

III. Methodology

We use quantitative operationalizations of the semantic features (semantic field, concept vagueness, salience and affect) as the predictors in a linear regression analysis. First, concerning the **semantic field** of the concepts, we rely on the division into chapters that is used in the dictionary. However, preliminary analyses indicated that a more fine-grained subdivision is necessary for the chapter about *the physical and abstract world* and for the chapter concerning a person’s *personality and feelings*. For these chapters, we use subsections in the dictionary to divide both chapters into two parts: *the physical world* versus *the abstract world* on the one hand, and *behavior* versus *feelings and intellect*, on the other hand. As a result, the predictor SEMANTIC FIELD has six levels: *the abstract world* (N = 115, e.g. DUIM, MAAT VAN 2.5 CM ‘~ an inch’), *the physical world* (N = 128, e.g. GELUID VAN NADEREND ONWEER ‘the sound of a storm that is approaching’), *the human body* (N = 180, e.g. NEUSGATEN ‘nostrils’), *family and sexuality* (N = 119, e.g. PEETOOM ‘godfather’), *behavior* (N = 93, e.g. HAASTIG ‘hasty, hurried’) and *feelings and intellect* (N = 224, e.g. SMALEN ‘to scorn’).

Formulating expectations concerning the influence of semantic field on lexical geographical variation is not easy, because different factors may play a role For instance, if more variation were to be found in *the physical world*, this could be explained by the fact that this semantic field contains a lot of concepts relating to the weather. Such concepts (like types of rain) are rather vague: they probably have fuzzy boundaries. However, it could also be argued that the background of the respondents may explain the number of variants that occurs for certain concepts. For example, a lot of lexical variants are available for weather concepts, especially for concepts regarding bad weather, because the respondents of the questionnaires on which the dictionary is based, were usually farmers, who were economically dependent on the weather [11, pp. V–VI].

Second, with regard to **vagueness**, we expect vaguer concepts, with fuzzy boundaries, to show more lexical geographical variation. We model the vagueness of a concept with the same method that was used in the pilot study. More specifically, we calculate the *lack of uniqueness* of the concept. Lack of uniqueness is defined as the number of lexical types per concept that occur for other concepts as well. The concept BEGRAVEN (‘to bury’), for instance, has a low value (0) for lack of uniqueness. KOUD, MISTIG EN SOMBER WEER (‘cold, misty, miserable weather’) has a high value (133). The latter concept seems to be relatively vague with regard to other bad weather concepts. The lexical item *dompetig weer*, for instance, is used to refer to both KOUD, MISTIG EN SOMBER WEER and to BENAUWD EN VOCHTIG WEER (‘damp, oppressive weather’). The lexical item *schuiverig* occurs for two other concepts: DRUILERIG EN KOUD WEER (‘dull, cold weather’) and GUUR, KIL EN SCHRAAL WEER (‘bleak, chilly weather’).

The third semantic feature that is included in the analysis pertains to the degree of salience of the concept. We expect less salient concepts to show more lexical geographical variation. We calculate the **lack of salience** of a concept with several measures.

First, we take into account the *relative number of multi-word expressions* (MWE’s) that occur per concept. This predictor was shown to have a significant effect on lexical geographical variation in the pilot study. The rationale for using this operationalization is two-fold. On the one hand, it relates to the basic-level hypothesis [12], which indicates that well-known, psychologically more entrenched concepts are generally expressed with shorter names. On the other hand, the data set also contains expressions that seem to have been chosen because the respondents were not familiar with either the concept itself or with the dialect name for the concept. A periphrastic response can, for example, be found in the data base for the concept VOORDE (‘ford’). Most respondents choose dialect words that are etymologically related to the standard

Dutch noun *voorde* or to the verb *waden* ('to wade through'), like *voerd* or *waajplaats*. However, a few respondents choose a periphrastic construction like *en stuk ondiep* (litt. 'a shallow part') or *ondepe plaatsj* (litt. 'shallow place'), which seems to indicate that they are not as familiar with the dialect name for the concept. We also investigated whether the proportion of periphrastic multi-word constructions can serve as an operationalization of lack of salience, but this factor did not reach significance in the regression model.

Second, the *relative number of places without a response* per concept in our data set is also used as an operationalization of lack of salience. This variable was used in the pilot study as well. It is based on the assumption that a high number of places without a response for a specific concept in the data base indicates that the respondents of the questionnaire do not know the dialect name for the concept. Note that this interpretation assumes that the questionnaire was distributed *systematically* across the *entire* language area.

The last operationalization of the lack of salience of a concept is based on the prevalence value of the lemma that is used in the dictionary to describe the concept. To operationalize *lack of prevalence*, we rely on data that was collected recently by [13]. In this study, which uses data from a large-scale online lexical decision experiment, word prevalence is defined as "the proportion of a population knowing a particular word" [13, p. 5]. However, since separate prevalence values were collected for Belgium and the Netherlands, while our data were collected in both countries, we use the minimum of the Belgian and Dutch prevalence score for the lemmas. We subtract this number from 1 to arrive at an approximation of the lack of prevalence of a certain concept. However, the fact that these data are much more recent than the dialect data may influence the results to a certain degree. A concept like SLIB, RIVIERBODEM ('silt'), for instance, has a rather high score for lack of prevalence (0.55), although this concept was probably relatively salient for the rural respondents of the questionnaires.

The fourth semantic feature in the analysis is **affect**. We use a different operationalization of affect than was used in the pilot study. On the one hand, we initially aimed to include a ternary predictor, that distinguishes between concepts with negative affect, concepts with positive affect and neutral concepts, while the pilot study used a numeric variable that modeled the degree to which a concept is prone to negative affect. However, since only few positive concepts (N = 32) occur in the data base, we rely on a concept's general sensitivity to affect instead. Second, a small-scale survey was used in the pilot study to determine how sensitive the human body concepts were to negative affect. However, in this paper, we use a binary predictor of affect sensitivity instead. More specifically, we coded each concept in the data base manually for its sensitivity to negative or positive affect (*sensitive*), or lack thereof (*neutral*). Overall, the data set contains 307 concepts that are sensitive to affect and 552 neutral concepts.

The response variable, **onomasiological heterogeneity**, takes two aspects of lexical geographical variation into account. On the one hand, we calculate the *diversity* of a concept as the number of word types that occurs in the dataset per concept. This factor models the difference between, for example, the concept BLOED ('blood'), for which only one lexeme (*bloed*) is found in the data base, and the concept LUIEREN ('to be lazy'), which occurs with 27 different word types in the data set. On the other hand, we also take into account *geographical scatter*. More specifically, we model the degree to which each concept and the variants per concept are scattered in a heterogeneous way across geographical space. Onomasiological heterogeneity is subsequently calculated as the logarithm of the product of diversity and scatter.

IV. RESULTS

Table I displays the output of the linear regression model. The model was built using a manual backwards selection procedure. With an Adjusted R^2 value of 0.5128, which indicates that about 51% of the variation in the data is explained by the model, the model performs relatively well. One outlier was found, but because leaving out this observation does not influence the model strongly, we discuss the model that includes the outlier in this section. Furthermore, we also built a model with interaction effects. Six interaction effects reached significance in this model (two interactions with SEMFIELD (LACK.OF.SALIENCE.relative.nr.mwe and LACK.OF.SALIENCE.relative.nr.missing.places) and four interactions with VAGUENESS (SEMFIELD, LACK.OF.SALIENCE.relative.nr.mwe, LACK.OF.SALIENCE.relative.nr.missing.places and AFFECT). However, as the interaction model only performs slightly better (Adjusted R^2 is 0.5624) than the main effects-only model and as the interactions don't contribute much to the interpretation of the main effects, we discuss the model without interactions below.

The model reveals that significant differences between semantic fields (**SEMFIELD**) occur. More specifically, we find that the chance of finding a large amount of lexical geographical variation is the highest in the semantic field of *the abstract world* (the reference level). The odds of finding less onomasiological heterogeneity are significantly smaller in the semantic fields *the human body*, *behavior* and *feelings and*

TABLE I. OUTPUT OF THE LINEAR REGRESSION MODEL: MAIN EFFECTS-ONLY MODEL

predictor	**estimate**	**p value**	
(Intercept)	3.027232	< 2e-16	***
SEMFIELD *the physical world*	-0.145553	0.333683	
SEMFIELD *the human body*	-0.423317	0.001695	**
SEMFIELD *family and sexuality*	-0.222910	0.139046	
SEMFIELD *behavior*	-0.567554	0.002642	**
SEMFIELD *feelings and intellect*	-0.635015	0.000170	***
LACK.OF.SALIENCE.relative.nr.mwe	2.689745	< 2e-16	***
LACK.OF.SALIENCE.relative.nr.missing.places	0.022385	0.000222	***
VAGUENESS	0.039539	< 2e-16	***
AFFECT *sensitive*	0.567249	2.61e-09	***

intellect.

Second, two predictors related to salience reach significance in the model, viz. the relative proportion of MWE's per concept and the relative number of places without a response for the concept. The estimates indicate that both predictors have the expected effect: the higher the proportion of MWE's and the higher the relative number of missing places (and, thus, the higher the lack of salience of the concept), the more likely it is that a larger amount of variation is found in the data.

The third predictor that reaches significance in the linear model pertains to the vagueness of a concept. More specifically, the model shows that, as expected, the chance of finding more onomasiological heterogeneity is higher for concepts with a high degree of lexical non-uniqueness.

Finally, we also found evidence for the importance of affect. More specifically, the model indicates that the chance of finding more lexical geographical variation in the data is higher for concepts that are sensitive to affect than for neutral concepts.

V. Discussion and conclusion

In this paper, we used a quantitative operationalization of the amount of lexical geographical variation that occurs for a concept in the dictionary of Limburgish dialects, to determine the influence of semantic concept features on this variable. Our analysis showed that the semantic features that were distinguished behave as expected.

First, differences between semantic fields occur: some semantic fields are less prone to variation than others. More specifically, we found that concepts from the field of *the abstract world* are significantly more likely than concepts relating to *the human body*, *behavior* or *feelings and intellect* to show a high degree of onomasiological heterogeneity if all other predictors are taken into account. A first explanation for this factor has to do with the distribution of the data. In particular, the concepts of *the abstract world* are generally rather salient and not vague, while the semantic fields of *behavior* and *feelings and intellect* contain a relatively large proportion of concepts with high values for the lack of salience and vagueness predictors. For VAGUENESS, for instance the median is 5 in *the abstract world*, but it is 10 in the field of *behavior* and 26 in *feelings and intellect*. Follow-up studies should consider building separate models for each of the semantic fields to determine the relative impact of the predictors in each semantic field.

A possible explanation for the finding that less variation occurs in the field of *the human body* than in *the abstract world*, is that concept concreteness may also have an influence on lexical geographical variation. The field of *the human body* seems to contain more concrete concepts (e.g. RUGGENGRAAT 'spinal column') than the semantic field of *the abstract world* (e.g. NAUW, ENG 'narrow'). However, further research is necessary to confirm that concept concreteness has an effect on lexical geographical variation.

Second, we found that the degree of salience of a concept (expressed by the relative number of multi-word expressions per concept and by the relative number of missing places per concept) also has the expected effect on lexical geographical variation: less salient concepts show more variation. However, *lack of prevalence* did not reach significance in the model. This may have to do with the fact that the prevalence data were collected much more recently than the dialect data. More specifically, it may be the case that the degree of prevalence of certain concepts has changed over time. Moreover, follow-up research should aim to include a profile-based calculation of lack of prevalence, which takes into account the prevalence values of all the lexical items that occur per concept

Aside from lack of salience, a second prototype-related concept feature, viz. vagueness, reached significance in the linear model as well. This confirms that concept vagueness plays a role in other semantic fields than *the human body*. Like in the pilot study, vaguer concepts, that are not clearly demarcated from related concepts, are more likely to show a higher amount lexical geographical variation.

Finally, the model indicated that a concept's susceptibility to affect has a significant influence on the amount of variation that is likely to occur. Concepts that are more sensitive to affect are more likely to show a higher amount of lexical geographical variation. However, as this predictor was coded manually, follow-up research should include a more objective operationalization of a concept's proneness to affect.

The Adjusted R^2 values of the linear models that were built in the pilot study were higher (about 0.6) than the Adjusted R^2 value (about 0.5) of this study. A possible explanation for this difference is that the present study does not contain all the quantifications of lack of salience and affect that reached significance in the pilot study. In particular, in the pilot study, lack of salience and negative affect were measured by means of small-scale surveys, in which respondents were asked to evaluate, on a scale of one to five, how likely it is that people are less familiar with the concepts and how prone the concepts are to negative affect. While we introduced two measures (namely an operationalization of prevalence and the predictor AFFECT, which was coded manually) to cope with the absence of the information on which the pilot study relied, it may be the case that these measures do not reach the same quality as explanatory factors of lexical geographical variation. However, as we included more than one semantic field in the analysis, while the pilot study focused on *the human body* alone, the data used in this study are also inherently more prone to variation. Ideally, follow-up research will investigate the influence of the survey-based measures of lack of salience and affect in other semantic fields than *the human body*, and more generally, the impact of semantic field on the results.

Some further lines of investigation can be envisaged. First, while this study uses the *product* of diversity (the number of lexical types per concept) and geographical scatter (the degree to which the variants are scattered in a homogeneous way across the dialect area) as the response variable of a linear regression analysis, the influence of concept characteristics should also be determined on both components of the response variable separately. Second, expansions to other dialect and

language areas will help to further establish the consistency of the impact of semantic features on lexical variation.

Overall, using a quantitative methodology, we were able to confirm that the amount of variation that is found in lexical data can, aside from by lectal and geographical factors, also be influenced by semantic concept features. Furthermore, we provided evidence for the importance of these types of features in other semantic fields than *the human body*. While differences between semantic fields occur, we found that a traditional feature of lexical variation (namely the sensitivity of a concept to affect) and two prototype-theoretical features (namely the degree of salience and the vagueness of a concept) significantly influence the variation found in dialect data.

REFERENCES

[1] K. Franco, D. Geeraerts and D. Speelman, "Lexical variation in a dialect area: influential features", in preparation.

[2] J. Nerbonne and P. Kleiweg, "Lexical Distance in LAMSAS," Comput. Hum., vol. 37, no. 3, pp. 339–357, 2003.

[3] J. Séguy, "La relation entre la distance spatiale et la distance lexicale," Rev. Linguist. Rom., vol. 35, pp. 335–357, 1971.

[4] M. Wieling, J. Nerbonne, and R. H. Baayen, "Quantitative social dialectology: Explaining linguistic variation geographically and socially," PLoS One, vol. 6, no. 9, p. e23613, Jan. 2011.

[5] D. Geeraerts and D. Speelman, "Heterodox concept features and onomasiological heterogeneity in dialects," in Advances in Cognitive Sociolinguistics, D. Geeraerts, G. Kristiansen, and Y. Peirsman, Eds. Berlin/New York: De Gruyter Mouton, 2010, pp. 23–40.

[6] D. Speelman and D. Geeraerts, "The role of concept characteristics in lexical dialectometry," Int. J. Humanit. Arts Comput., vol. 2, pp. 221–242, 2009.

[7] K. Allan and K. Burridge, "Euphemism, dysphemism, and cross-varietal synonymy," La Trobe Work. Pap. Linguist., vol. 1, 1988.

[8] K. Allan and K. Burridge, Forbidden Words: Taboo and the Censoring of Language. Cambridge: Cambridge University Press, 2006.

[9] "Woordenboek van de Limburgse Dialecten." Van Gorcum/Gopher, Assen/Groningen, 1983-2008.

[10] L. Vandermeeren, Het Bjêvels. Beverlo, 1995.

[11] J. Kruijsen, Woordenboek van de Limburgse Dialecten. III.4.4: De stoffelijke en abstracte wereld. Groningen: Gopher Publishers, 2004.

[12] B. Berlin, D. Breedlove, and P. Raven, "General principles of classification and nomenclature in folk biology," Am. Anthropol., vol. 75, pp. 214–242, 1973.

[13] E. Keuleers, M. Stevens, P. Mandera, and M. Brysbaert, "Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment," Q. J. Exp. Psychol., 2015.